W478–W483 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 9 May 2012*
doi:10.1093/nar/gks402

GeneCodis3: a non-redundant and modular enrichment analysis tool for functional genomics

Daniel Tabas-Madrid, Ruben Nogales-Cadenas and Alberto Pascual-Montano*

Functional Bioinformatics Group, National Center for Biotechnology (CNB-CSIC), Madrid, Spain

Received February 13, 2012; Revised April 10, 2012; Accepted April 18, 2012

ABSTRACT

Since its first release in 2007, GeneCodis has become a valuable tool to functionally interpret results from experimental techniques in genomics. This web-based application integrates different sources of information to finding groups of genes with similar biological meaning. This process, known as enrichment analysis, is essential in the interpretation of high-throughput experiments. The frequent feedbacks and the natural evolution of genomics and bioinformatics have allowed the growth of the tool and the development of this third release. In this version, a special effort has been made to remove noisy and redundant output from the enrichment results with the inclusion of a recently reported algorithm that summarizes significantly enriched terms and generates functionally coherent modules of genes and terms. A new comparative analysis has been added to allow the differential analysis of gene sets. To expand the scope of the application, new sources of biological information have been included, such as genetic diseases, drugs–genes interactions and Pubmed information among others. Finally, the graphic section has been renewed with the inclusion of new interactive graphics and filtering options. The application is freely available at http://genecodis.cnb.csic.es.

INTRODUCTION

Next-generation sequencing and other high-throughput experimental techniques such as DNA microarrays are revolutionary technologies for advanced biomedical research. Instead of studying individual genes or proteins, these techniques tackle the study of different experimental conditions from a global perspective. A large list of genes or proteins is usually resulted after processing the massive amount of information generated by these techniques. The interpretation of these lists in a biological context is the next logical step.

Several data mining techniques have been developed for a better understanding of the experiments. One of the most popular methods consists in finding biological annotations statistically enriched in lists of genes or proteins compared with a reference set, which is often the entire genome or the complete list of genes in a study. Tools that translate large lists of genes or proteins into related terms from Gene Ontology (GO) (1) such as Onto-Express (2), DAVID (3), FATIGO+ (4) or ProfCom (5) are now routinely used. A good compendium of available tools can be found at http://www.geneontology.org/GO.tools.shtml (accessed April 30, 2012).

According to a recent review (6), current functional enrichment algorithms can be classified into three groups: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) and modular enrichment analysis (MEA). SEA is probably the most typical way to make an enrichment analysis (7,8). This strategy looks for individual annotations related to genes or proteins in different biological databases and evaluates the statistical significance of each individual term. This is commonly made by counting the number of genes from the input and reference lists containing the term and computing a statistical test, usually χ^2 or Fisher's test and using Binomial or Hypergeometric distributions. The computed *P*-value represents the significance of the association between the term and the input list of molecules.

GSEA (9) works in the same way as SEA but taking into account the whole list of genes instead of analyzing the most relevant ones. Its main goal is to avoid the arbitrary cutoff in expression level. The main idea is to know if the set of genes related with a specific term tends to concentrate at the top (the most overexpressed) or at the bottom (the most inhibited) of the list. There are few available tools for GSEA like FatiScan (10) or GO-Mapper (11).

Although these strategies have been demonstrated to be useful, the lack of term–term relationships in their analyses is an important drawback. The combinations of terms extend our understanding of biological events associated with a given experimental system. It provides a more specific meaning while increasing the quality of results by detecting additional significant terms.

*To whom correspondence should be addressed. Tel: +34 915854617; Fax: +34 913720112; Email: pascual@cnb.csic.es

As shown in (12), there are biological terms not detected as statistically relevant when tested individually, which turn out to be significant when appearing together with others. MEA methods avoid this problem and provide a more complete functional analysis. Two examples of MEA tools are DAVID, that does clustering of singular enrichment results and ProfCom, that takes into account different kinds of logic relationships between the terms (AND, OR, NOT).

In 2007, we introduced GeneCodis (12,13), a tool for singular and modular enrichment analysis that integrates information of diverse nature (e.g. functional, regulatory or structural) by looking for frequent patterns in the space of annotations and computing their statistical relevance. This integrative capacity sheds light on different aspects of the same information and provides a more accurate interpretation of the data.

In this work, we present a new release of GeneCodis with several novel features. New types of annotations now expand its analysis capacity. For example, Pubmed Ids, pathways from Panther (14) or drug target genes from PharmGKB (15). Exome analysis is also possible by the inclusion of exon ids from Ensembl (16). Another significant novelty consists in comparing different lists of genes or proteins (17,18), focused in finding shared and divergent information related with different experimental conditions. GeneCodis3 also integrates a new post-analysis algorithm to remove the noise and the intrinsic redundancy of the biological annotations. Finally, a new graphical section has been implemented, with new interactive visualizations and the option to filter and highlight terms.

NEW FEATURES

The standard workflow of the tool has not been drastically modified to keep the same logic and consistency with the previous version. The analysis is made by pasting or loading a list of genes and selecting the organism and the categories of annotations. The job is sent to our servers, and once the results are ready, all visualizations are displayed. MEA and SEA are always simultaneously executed for every submission. Keeping the web browser opened while executing the job is not necessary. Results can be retrieved by keeping the active URL or by introducing an email address to receive a notification when the job is ready. The new features, explained more deeply below, are intended to facilitate the interpretation of the results and to make the system more self-contained. Table 1 summarizes the differences between previous and current versions.

Non-redundant reciprocal linkage of genes and terms

Enrichment analysis techniques are very useful to extract biological knowledge from a set of genes or proteins, but at the same time, they frequently generate a long list of significant terms, even for a very small set of molecules. The problem is worse when different annotations are used because the size of the problem increases significantly.

Most of these large sets of results suffer from the redundancy of the biological terms that are repeated in many different annotation resources. For example, GO, KEGG (19) and others have different terms to express the same event (e.g. Glycolysis GO:0006096 in GO and Glycolysis/Glucogenesis 00010 pathway in KEGG). The GO hierarchy also affects the results since some similar terms are used to express the same reality (e.g. GO:0065007 biological regulation and GO:0050789 regulation of biological process). Additionally, the bias in the annotation space also affects the enrichment since there are frequent general terms that do not provide very useful information, such as the GO:0007049 cell cycle term.

A recently proposed computational method, called GeneTerm Linker or GTLinker for short (20), summarizes the enrichment results by finding significant and coherent collections of genes and terms. This approach executes several filtering and clustering steps to eliminate redundant and non-informative terms and produces genes and annotations grouped in modules (metagroups). These metagroups, which are supposed to be functionally coherent, absorb the redundancy of the original significant terms and are ranked by their significance and coherence.

We have implemented this method in GeneCodis as an optional post-analysis of the results. The output contains a table with the cluster of genes and annotations as well as their statistical significance. Details can also be displayed to show the original sets of the enrichment analysis, and the similarity coefficient of the resulting group. Currently, only *Homo sapiens* and *Saccharomyces cerevisiae* are supported, but the rest of organisms will be incrementally added.

Comparative analysis

It is very common in high-throughput experiments to work with different conditions that finally derive in different groups of genes. In these cases, the challenge consists in interpreting the functional differences of the groups. Executing an enrichment analysis for each set of genes and manually comparing them is the traditional way to do it. This time-consuming task could be difficult when working with large lists.

We have now included a comparative analysis tool that allows the submission of two different lists at the same time and carries out a simultaneous modular and singular enrichment analyses for each one. The most valuable information when comparing two lists is the common and divergent information. Three additional analyses are performed for the intersection of the two sets as well as the original ones without the intersecting area. These operations are done over the space of genes instead of the annotations to guarantee the statistical significance of the results.

The results are shown in a similar way to the results of a standard analysis. For each category, an interactive Venn diagram with the number of genes of each group is

Table 1. Comparative table between previous and current versions of GeneCodis

Features	GeneCodis2	GeneCodis3
Singular enrichment analysis	✓	✓
Modular enrichment analysis	✓	✓
Comparative analysis		✓
Redundancy removal (GTLinker)		✓
Functional annotations	✓	✓
Regulatory annotations	✓	✓
Pharmacological annotations		✓
Disease annotations		✓
Panther pathways annotations		✓
Pubmed annotations		✓
Graphics section	✓	✓
Interactive visualization		✓
Tag cloud of terms		✓
Computing cluster support	✓	✓
Multithreading support		✓

displayed. When clicking on each group of the diagram, a table with its corresponding results is opened. Results for the original sets are also included.

New sources of annotations

Due to the highly distributed nature of biological information, one of the most important tasks of GeneCodis consists in the integration of all the data in a unique local resource to allow an assortment of analyses. We have used Biomart (21) to automate these data accession and management tasks.

GeneCodis already contains an extensive amount of information from different resources. For example, GO, KEGG pathways, structural information via the InterPro protein domains (22) and regulatory information from microRNAs targets from mirBase (23). Transcription factors (TFs) were extracted from curated sources. For human, rat and mouse, TFs were extracted from the MsigDB database (9), mainly based on TransFac (24). In the case of yeast, TFs were obtained from the experimental data contained in YEASTRACT database (25).

This release supports gene-related diseases from the online Mendelian inheritance in man (OMIM) database (26), which contains complete information about the known mendelian disorders. Information from PharmGKB is included to measure the effect of genetic activity in response to drugs. This annotation increases its value when considered together with OMIM. Panther is another ontology-based database of pathways (mainly signaling pathways) that, jointly with KEGG, provides an excellent resource to know the map of interactions of a given gene list. The inclusion of biomedical literature data from Pubmed identifiers (http://www.ncbi.nlm.nih.gov/pubmed/, accessed April 30, 2012) facilitates the finding of articles related with a group of genes. Finally, in consonance with the recent increase of exome resequencing studies, GeneCodis accepts exons identifiers from Ensembl to find the significant terms that are overrepresented in the genes composed by any of the input exons.

User interface

When analyzing big sets of genes, results are usually too large for visual analysis. Graphics such as pie or column charts are very useful to quickly detect the most significant combinations of terms. New interactive graphics have been integrated in this version, as well as filters to select the groups of annotations, either by name or by number of genes.

GeneCodis also includes a new tag cloud that displays the 30 most significant terms of the results. The tag's sizes vary according to the number of supporting genes and are also linked to their corresponding databases. This graphical representation allows a rapid interpretation of the most relevant content of the analysis.

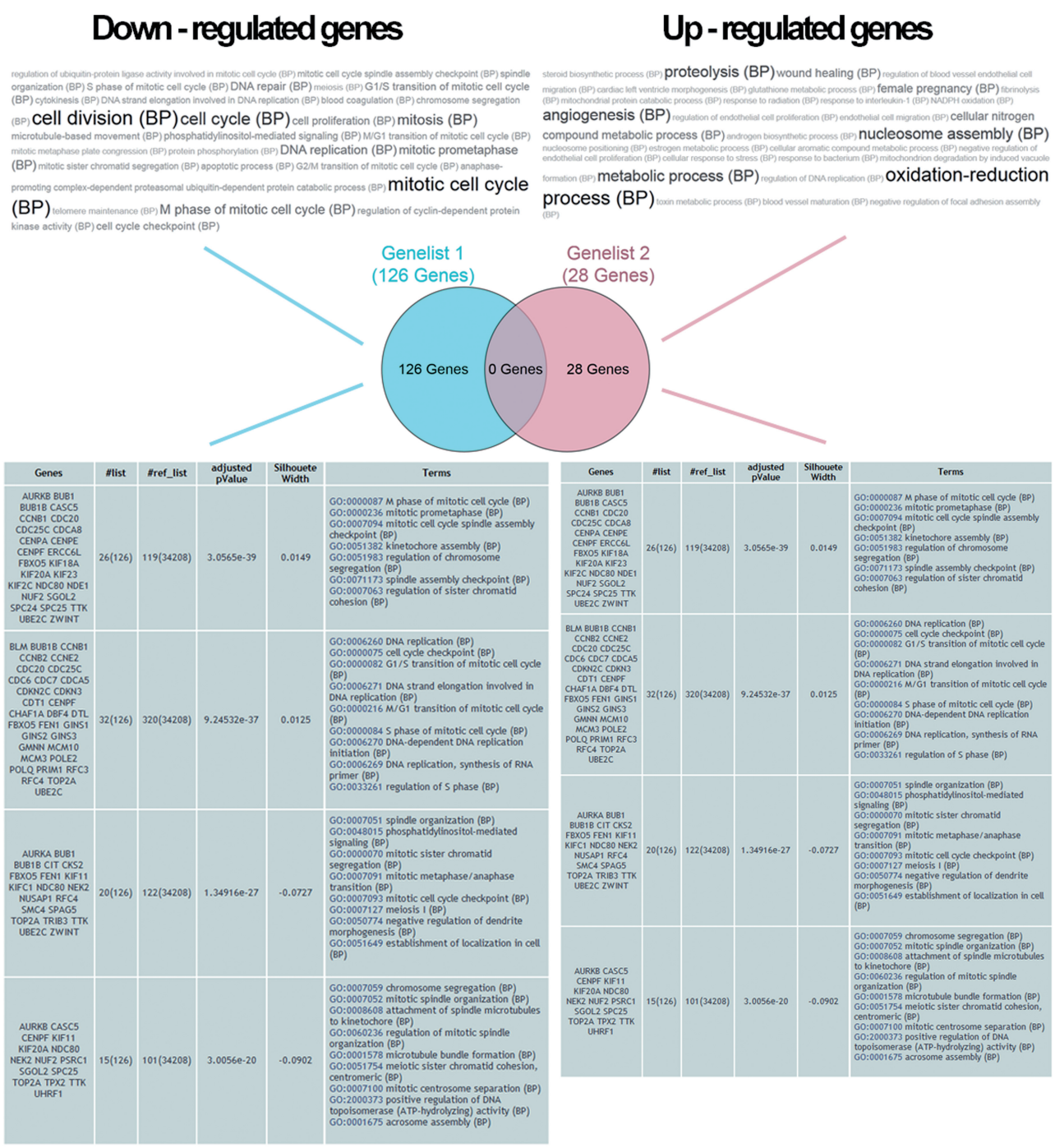
Like in the previous release, GeneCodis supports programmatically access through the SOAP Web Services technology. With a few lines of code, GeneCodis can be integrated in any kind of pipeline, regardless of the programming language or platform.

GENECODIS USE CASE AND OUTPUT DESCRIPTION

We demonstrate the utility of GeneCodis by analyzing the differentially expressed genes of human lung WI-38 fibroblasts upon exposure to the ultimate carcinogen benzo[*a*]pyrene diol epoxide (BPDE) (17). In this study, normal human WI-38 lung fibroblasts were exposed to three different concentrations of BPDE, and whole-genome oligonucleotide microarrays were used to measure changes in gene expression. WI-38 cells were exposed to 0.1, 0.5 or 1 μM BPDE for 24 h which resulted in 384, 972 and 837 differentially expressed genes, respectively. Results of this study are summarized in Figure 1 of (17).

Using the same list of genes that were reported as expressed in response to all doses of BPDE, we repeated the analysis with GeneCodis using the Comparative option, which is very convenient for experimental designs involving more than one condition. Results are shown in Figure 1. As input we submitted two lists of genes, the organism (human in this case) and the annotations. To reproduce the results in the article, we only used GO Biological Process.

The central part of Figure 1 shows a Venn diagram containing the number of genes in each list after annotation. Genes with no annotations are removed and not considered in the statistics. The top of the figure shows the cloud of tags that contains significant enriched terms in each analysis. A simple visual inspection already provides information about the functional implication of the gene sets. As explained in (17), downregulated genes are involved in mitosis, cell cycle, spindle organization and biogenesis as well as DNA replication and DNA repair. Similarly, biological processes enriched in upregulated genes affected by all doses of BPDE are highlighted in the right cloud of tags and show the involvement of all these genes in the nucleosome assembly, cell proliferation and extracellular processes.



Genes	#list	#ref_list	adjusted pValue	Silhouete Width	Terms
AURKB BUB1 BUB1B CASC5 CCNB1 CDC20 CDC25C CDCA8 CENPA CENPE CENPF ERCC6L FBXO5 KIF18A KIF20A KIF23 KIF2C NDC80 NDE1 NUF2 SGOL2 SPC24 SPC25 TTK UBE2C ZWINT	26(126)	119(34208)	3.0565e-39	0.0149	GO:0000087 M phase of mitotic cell cycle (BP) GO:0000236 mitotic prometaphase (BP) GO:0007094 mitotic cell cycle spindle assembly checkpoint (BP) GO:0051382 kinetochore assembly (BP) GO:0051983 regulation of chromosome segregation (BP) GO:0071173 spindle assembly checkpoint (BP) GO:0007063 regulation of sister chromatid cohesion (BP)
BLM BUB1B CCNB1 CCNB2 CCNE2 CDC20 CDC25C CDC6 CDC7 CDCA5 CDKN2C CDKN3 CDT1 CENPF CHAF1A DBF4 DTL FBXO5 FEN1 GINS1 GINS2 GINS3 GMNN MCM10 MCM3 POLE2 POLQ PRIM1 RFC3 RFC4 TOP2A UBE2C	32(126)	320(34208)	9.24532e-37	0.0125	GO:0006260 DNA replication (BP) GO:0000075 cell cycle checkpoint (BP) GO:0000082 G1/S transition of mitotic cell cycle (BP) GO:0006271 DNA strand elongation involved in DNA replication (BP) GO:0000216 M/G1 transition of mitotic cell cycle (BP) GO:0000084 S phase of mitotic cell cycle (BP) GO:0006270 DNA-dependent DNA replication initiation (BP) GO:0006269 DNA replication, synthesis of RNA primer (BP) GO:0033261 regulation of S phase (BP)
AURKA BUB1 BUB1B CIT CKS2 FBXO5 FEN1 KIF11 KIFC1 NDC80 NEK2 NUSAP1 RFC4 SMC4 SPAG5 TOP2A TRIB3 TTK UBE2C ZWINT	20(126)	122(34208)	1.34916e-27	-0.0727	GO:0007051 spindle organization (BP) GO:0048015 phosphatidylinositol-mediated signaling (BP) GO:0000070 mitotic sister chromatid segregation (BP) GO:0007091 mitotic metaphase/anaphase transition (BP) GO:0007093 mitotic cell cycle checkpoint (BP) GO:0007127 meiosis I (BP) GO:0050774 negative regulation of dendrite morphogenesis (BP) GO:0051649 establishment of localization in cell (BP)
AURKB CASC5 CENPF KIF11 KIF20A NDC80 NEK2 NUF2 PSRC1 SGOL2 SPC25 TOP2A TPX2 TTK UHRF1	15(126)	101(34208)	3.0056e-20	-0.0902	GO:0007059 chromosome segregation (BP) GO:0007052 mitotic spindle organization (BP) GO:0008608 attachment of spindle microtubules to kinetochore (BP) GO:0060236 regulation of mitotic spindle organization (BP) GO:0001578 microtubule bundle formation (BP) GO:0051754 meiotic sister chromatid cohesion, centromeric (BP) GO:0007100 mitotic centrosome separation (BP) GO:2000373 positive regulation of DNA topoisomerase (ATP-hydrolyzing) activity (BP) GO:0001675 acrosome assembly (BP)

Genes	#list	#ref_list	adjusted pValue	Silhouete Width	Terms
AURKB BUB1 BUB1B CASC5 CCNB1 CDC20 CDC25C CDCA8 CENPA CENPE CENPF ERCC6L FBXO5 KIF18A KIF20A KIF23 KIF2C NDC80 NDE1 NUF2 SGOL2 SPC24 SPC25 TTK UBE2C ZWINT	26(126)	119(34208)	3.0565e-39	0.0149	GO:0000087 M phase of mitotic cell cycle (BP) GO:0000236 mitotic prometaphase (BP) GO:0007094 mitotic cell cycle spindle assembly checkpoint (BP) GO:0051382 kinetochore assembly (BP) GO:0051983 regulation of chromosome segregation (BP) GO:0071173 spindle assembly checkpoint (BP) GO:0007063 regulation of sister chromatid cohesion (BP)
BLM BUB1B CCNB1 CCNB2 CCNE2 CDC20 CDC25C CDC6 CDC7 CDCA5 CDKN2C CDKN3 CDT1 CENPF CHAF1A DBF4 DTL FBXO5 FEN1 GINS1 GINS2 GINS3 GMNN MCM10 MCM3 POLE2 POLQ PRIM1 RFC3 RFC4 TOP2A UBE2C	32(126)	320(34208)	9.24532e-37	0.0125	GO:0006260 DNA replication (BP) GO:0000075 cell cycle checkpoint (BP) GO:0000082 G1/S transition of mitotic cell cycle (BP) GO:0006271 DNA strand elongation involved in DNA replication (BP) GO:0000216 M/G1 transition of mitotic cell cycle (BP) GO:0000084 S phase of mitotic cell cycle (BP) GO:0006270 DNA-dependent DNA replication initiation (BP) GO:0006269 DNA replication, synthesis of RNA primer (BP) GO:0033261 regulation of S phase (BP)
AURKA BUB1 BUB1B CIT CKS2 FBXO5 FEN1 KIF11 KIFC1 NDC80 NEK2 NUSAP1 RFC4 SMC4 SPAG5 TOP2A TRIB3 TTK UBE2C ZWINT	20(126)	122(34208)	1.34916e-27	-0.0727	GO:0007051 spindle organization (BP) GO:0048015 phosphatidylinositol-mediated signaling (BP) GO:0000070 mitotic sister chromatid segregation (BP) GO:0007091 mitotic metaphase/anaphase transition (BP) GO:0007093 mitotic cell cycle checkpoint (BP) GO:0007127 meiosis I (BP) GO:0050774 negative regulation of dendrite morphogenesis (BP) GO:0051649 establishment of localization in cell (BP)
AURKB CASC5 CENPF KIF11 KIF20A NDC80 NEK2 NUF2 PSRC1 SGOL2 SPC25 TOP2A TPX2 TTK UHRF1	15(126)	101(34208)	3.0056e-20	-0.0902	GO:0007059 chromosome segregation (BP) GO:0007052 mitotic spindle organization (BP) GO:0008608 attachment of spindle microtubules to kinetochore (BP) GO:0060236 regulation of mitotic spindle organization (BP) GO:0001578 microtubule bundle formation (BP) GO:0051754 meiotic sister chromatid cohesion, centromeric (BP) GO:0007100 mitotic centrosome separation (BP) GO:2000373 positive regulation of DNA topoisomerase (ATP-hydrolyzing) activity (BP) GO:0001675 acrosome assembly (BP)

Figure 1. Collage of output generated by GeneCodis. The Venn diagram contains the number of actual genes in each list after annotation. Two tag cloud visualizations containing significant enriched terms in each analysis are shown on the top. The two tables on the bottom show the four most relevant metagroups of genes and terms generated after running the integrated GTLinker on each list to avoid redundancy.

The bottom of Figure 1 shows the metagroups generated after running the integrated GTLinker to avoid redundancy in the original enrichment results. In the case of repression, GeneCodis generated 203 statistical significant terms and GTLinker clustered them into a reduced set of 41 metagroups. The left table in Figure 1 shows the top four metagroups according to their significance (*P*-value). In a similar fashion, the table in the right shows the top four metagroups out of a total of eight metagroups that were generated from the 57 enrichment results of the induced genes. By exploring the GO terms in each metagroup, we can identify the main biological processes that are related to each condition. The coherence and succinctness of the information provided by the most significant metagroups, together with the comparative functionality and visualizations, have made the reproducibility of this study quite simple and automatic.

TECHNICAL DETAILS

GeneCodis core consists in a C++ code which includes a method to generate combination of frequent biological terms that are present in the gene sets (27). A Ruby wrapper processes all data and creates a script to run the core program, balancing the computational load by deciding which processor will execute the job. The Ruby wrapper has been improved from the previous versions by taking advantage of new releases of this language, making use of threads to re-implement potentially parallel parts of the program.

Client–server architecture has been used to decouple the front-end from the heavy computational code that runs on a different server. We use a cluster of six new-generation servers, each one containing two Quad-Core Intel Xeon 64 bits processors, that is able to handle a large number of simultaneous jobs. The web portal has been ported to Ruby on Rails, a well-tested and stable technology. As in the previous version, the whole functionality is also supported by web services to allow programmatic access to the tool.

This new version has been running for a few months, experiencing more than 1000 submissions. Extensive tests have been carried out in different web browsers using synthetic and real data sets for which the outcome of the software is known.

DISCUSSION

Since the launch of its second release, GeneCodis received more than 43 500 effective visits, each of them executing several jobs, which gives an indication of the popularity of the tool. One of the disadvantages of enrichment analysis and more in particular with the modular enrichment is the generation of hundreds of significant results loaded with an inherent redundancy that make interpretation very difficult. This is caused by the natural redundancy of the information across biological databases. In this new release, we have tackled this issue with the inclusion of a recently reported algorithm, which groups and filters the data extracted from the enrichment analysis. This is complemented with the addition of several new sources of information.

The comparative functionality also represents a qualitative improvement. It allows an easy comparison of gene sets in the same step by executing simultaneous analyses of their possible combinations. This is in concert with the common experimental designs in high-throughput experiments. The new graphical section also improves the way to visualize the results, allowing a direct interaction with the charts and tables. A special attention should be given to the cloud of tags, which produces a quick snapshot of the most prominent and significant enriched terms. Even if this visualization was not intended to summarize the information, it acts effectively as such.

The new functionality, extensively described in this article, together with the new visualization and computational infrastructure, makes GeneCodis a novel functional genomics tool compared not only with its previous versions but also with some of the other existing tools in the field.

FUNDING

Spanish Minister of Science and Innovation [BIO2010-17527]; Government of Madrid (CAM) [P2010/BMD-2305]; Juan de la Cierva research program (to R.N.-C.). Funding for open access charge: Spanish Minister of Science and Innovation [BIO2010-17527].

Conflict of interest statement. None declared.

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al*.; The Gene Ontology Consortium. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics*, **25**, 25–29.
2. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
3. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.
4. Al-Shahrour,F., Minguez,P., Tárraga,J., Medina,I., Alloza,E., Montaner,D. and Dopazo,J. (2007) FatiGO +: a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–W96.
5. Antonov,A.V., Schmidt,T., Wang,Y. and Mewes,H.W. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.
6. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.
7. Khatri,P. and Drăghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
8. Dopazo,J. (2006) Functional interpretation of microarray experiments. *Omics*, **10**, 398–410.
9. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al*. (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
10. Al-Shahrour,F., Arbiza,L., Dopazo,H., Huerta-Cepas,J., Mínguez,P., Montaner,D. and Dopazo,J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.
11. Smid,M. and Dorssers,L.C.J. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics*, **20**, 2618–2625.
12. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.
13. Nogales-Cadenas,R., Carmona-Saez,P., Vazquez,M., Vicente,C., Yang,X., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2009) GeneCodis: interpreting gene lists through enrichment analysis and integration of diverse biological information. *Nucleic Acids Res.*, **37**, W317–W322.
14. Mi,H., Dong,Q., Muruganujan,A., Gaudet,P., Lewis,S. and Thomas,P.D. (2010) PANTHER version 7: improved

phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.*, **38**, D204–D210.
15. McDonagh,E.M., Whirl-Carrillo,M., Garten,Y., Altman,R.B. and Klein,T.E. (2011) From pharmacogenomic knowledge acquisition to clinical applications: the PharmGKB as a clinical pharmacogenomic biomarker resource. *Biomark. Med.*, **5**, 795–806.
16. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Brent,S., Chen,Y., Clapham,P., Coates,G., Fairley,S., Fitzgerald,S. *et al.* (2011) Ensembl 2011. *Nucleic Acids Res.*, **39**, D800–D806.
17. Dreij,K., Rhrissorrakrai,K., Gunsalus,K.C., Geacintov,N.E. and Scicchitano,D.A. (2010) Benzo[a]pyrene diol epoxide stimulates an inflammatory response in normal human lung fibroblasts through a p53 and JNK mediated pathway. *Carcinogenesis*, **31**, 1149–1157.
18. Rezaul,K., Thumar,J.K., Lundgren,D.H., Eng,J.K., Claffey,K.P., Wilson,L. and Han,D.K. (2010) Differential protein expression profiles in estrogen receptor-positive and -negative breast cancer tissues using label-free quantitative proteomics. *Genes Cancer*, **1**, 251–271.
19. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2011) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
20. Fontanillo,C., Nogales-Cadenas,R., Pascual-Montano,A. and De Las Rivas,J. (2011) Functional analysis beyond enrichment: non-redundant reciprocal linkage of genes and biological terms. *PLoS One*, **6**, e24289.
21. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, November 11 (doi:10.1093/database/bar049; epub ahead of print).
22. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2011) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
23. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
24. Matys,V. (2003) TRANSFAC(R): transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, **31**, 374–378.
25. Abdulrehman,D., Monteiro,P.T., Teixeira,M.C., Mira,N.P., Lourenço,A.B., dos Santos,S.C., Cabrito,T.R., Francisco,A.P., Madeira,S.C., Aires,R.S. *et al.* (2011) YEASTRACT: providing a programmatic access to curated transcriptional regulatory associations in Saccharomyces cerevisiae through a web services interface. *Nucleic Acids Res.*, **39**, D136–D140.
26. Hamosh,A. (2002) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.*, **30**, 52–55.
27. Borgelt,C., Yang,X., Nogales-Cadenas,R., Carmona-Saez,P. and Pascual-Montano,A. (2011) Finding closed frequent item sets by intersecting transactions. In: *Proceedings of the 14th International Conference on Extending Database Technology - EDBT/ICDT '11*. ACM Press, New York, USA, p. 367.